Exploiting Similarities among Languages for Machine Translation

Mikolov et al. 2013

Gliederung

- Einführung
- Bestimmung einer Translation-Matrix
- Experimente und Baselines
- Evaluation
- Fazit

Einführung

- Wörterbücher und Phrasentabellen waren Grundlage vieler moderner Verfahren maschineller Übersetzung

Einführung

- Wörterbücher und Phrasentabellen waren Grundlage vieler moderner Verfahren maschineller Übersetzung

  → Automatische Verfahren zur Erweiterung und Korrektur dieser Wörterbücher sind wünschenswert

Einführung

- Wörterbücher und Phrasentabellen waren Grundlage vieler moderner Verfahren maschineller Übersetzung

  → Automatische Verfahren zur Erweiterung und Korrektur dieser Wörterbücher sind wünschenswert
- Idee:
- 1. Nutze besser verfügbare monolinguale Korpora und erzeuge damit monolinguale Modelle

Einführung

- Wörterbücher und Phrasentabellen waren Grundlage vieler moderner Verfahren maschineller Übersetzung

  → Automatische Verfahren zur Erweiterung und Korrektur dieser Wörterbücher sind wünschenswert
- Idee:

  1. Nutze besser verfügbare monolinguale Korpora und erzeuge damit monolinguale Modelle

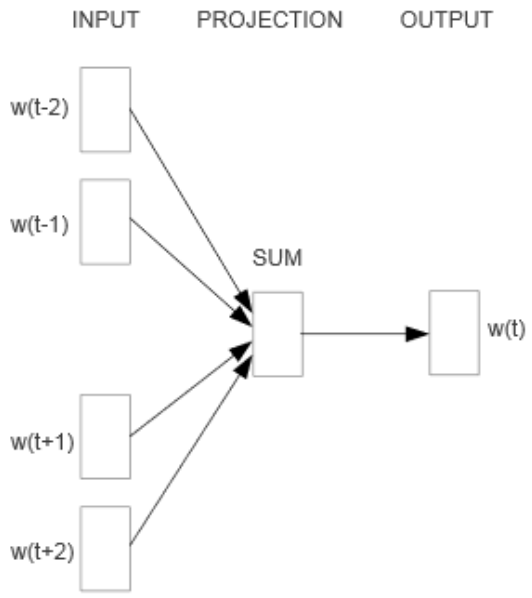
  2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden

Einführung

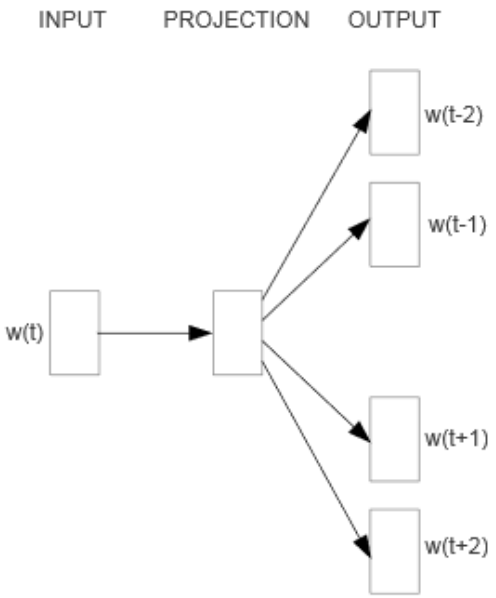
- 1. Nutze besser verfügbare monolinguale Korpora und erzeuge damit monolinguale Modelle

Einführung

- 1. Nutze besser verfügbare monolinguale Korpora und erzeuge damit monolinguale Modelle
  - Erstellung monolingualer Modelle wie gewohnt (Skip-gram/CBOW)



CBOW



Skip-gram

Einführung

- 1. Nutze besser verfügbare monolinguale Korpora und erzeuge damit monolinguale Modelle
  - Erstellung monolingualer Modelle wie gewohnt (Skip-gram/CBOW)
  - Parallele Implementierung

Einführung

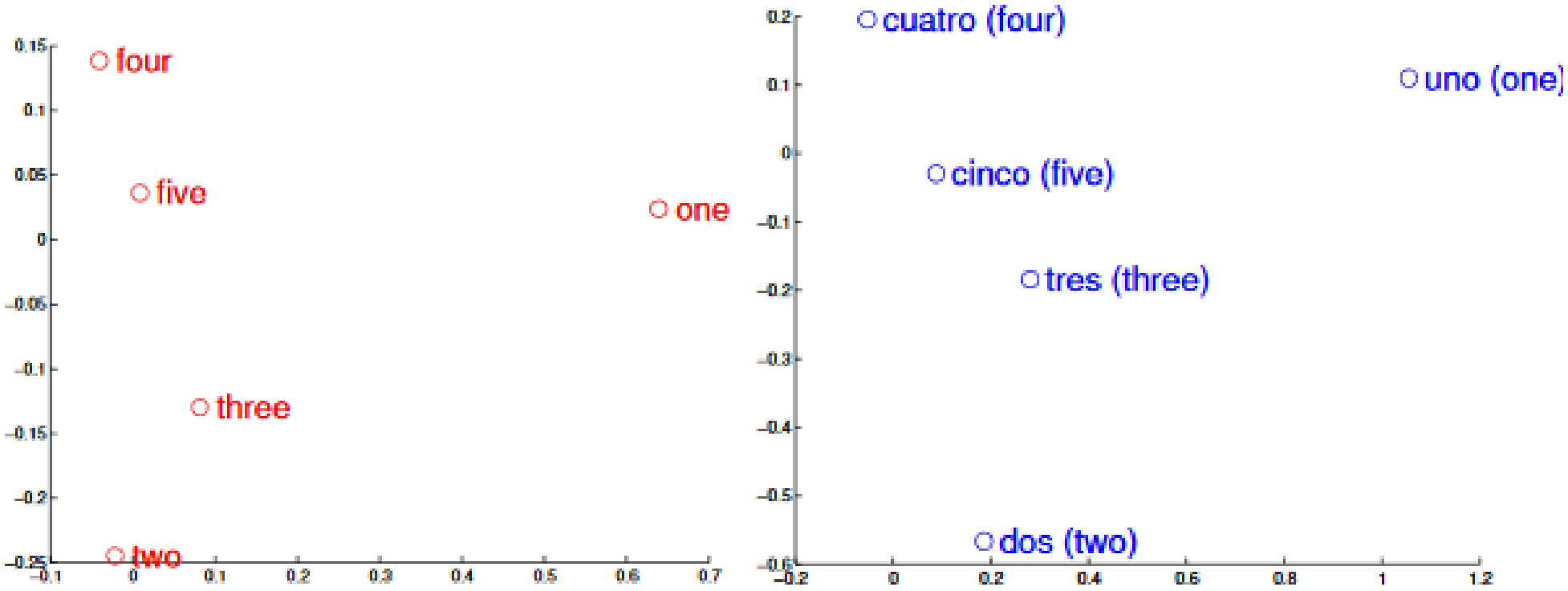
- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden

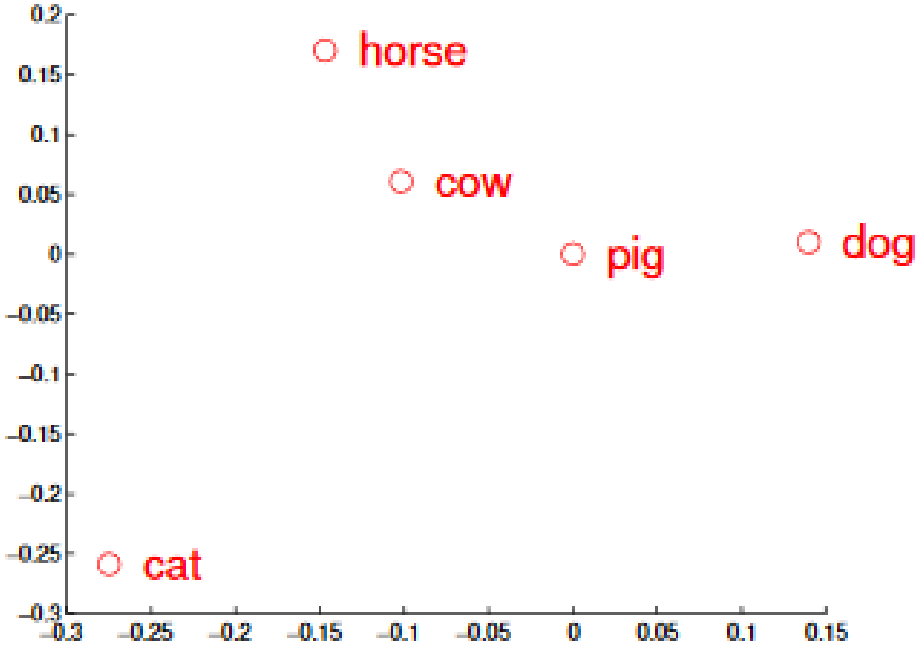


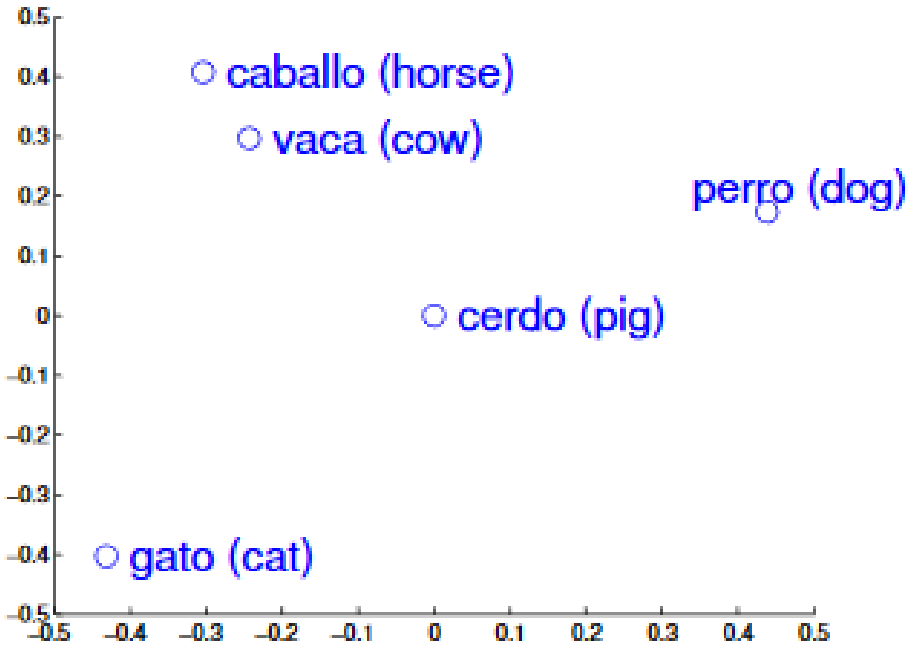
Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden





Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden

  → Wie kommt diese Beziehung zu Stande?

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden
- Sprachen existieren nicht losgelöst von der Realität, sondern basieren auf realen Konzepten, die für alle Sprachen gelten

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden
- Sprachen existieren nicht losgelöst von der Realität, sondern basieren auf realen Konzepten, die für alle Sprachen gelten

  Möglichkeit arithmetischer Operationen wie

  Friseuse – Frau + Mann ≈ Friseur

  zeigt, dass reale Gegebenheiten (wie Geschlecht, Größe, Zusammenhänge, etc.) in den Vektorräumen kodiert sind

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden
- Sprachen existieren nicht losgelöst von der Realität, sondern basieren auf realen Konzepten, die für alle Sprachen gelten

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden
- Sprachen existieren nicht losgelöst von der Realität, sondern basieren auf realen Konzepten, die für alle Sprachen gelten
- Wir suchen also eine Möglichkeit, die Vektorräume verschiedener Sprachen aufeinander zu mappen

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden
- Sprachen existieren nicht losgelöst von der Realität, sondern basieren auf realen Konzepten, die für alle Sprachen gelten
- Wir suchen also eine Möglichkeit, die Vektorräume verschiedener Sprachen aufeinander zu mappen
- Wie können wir Vektoren rotieren und skalieren?

Einführung

- 2. Nutze die weniger vorhandenen bilingualen Ressourcen, um ein Mapping zwischen beiden Modellen zu finden
- Sprachen existieren nicht losgelöst von der Realität, sondern basieren auf realen Konzepten, die für alle Sprachen gelten
- Wir suchen also eine Möglichkeit, die Vektorräume verschiedener Sprachen aufeinander zu mappen
- Wie können wir Vektoren rotieren und skalieren?

  → Wir bestimmen eine Matrix

Bestimmung einer Translation-Matrix

- Gegeben:
  - Menge an Wortpaaren (im Wörterbuch) $\{x_i, z_i\}_{i=1}^{n}$
  - 2 Vektorrepräsentationen pro Paar $x_i \in \mathbb{R}^{d_1}, z_i \in \mathbb{R}^{d_2}$

Bestimmung einer Translation-Matrix

- Gegeben:

  Menge an Wortpaaren (im Wörterbuch) $\{x_i, z_i\}_{i=1}^{n}$

  2 Vektorrepräsentationen pro Paar $x_i \in \mathbb{R}^{d_1}, z_i \in \mathbb{R}^{d_2}$

- Gesucht: Transformationsmatrix *W*

  - sodass: $\sum_{i=1}^{n} \|Wx_i - z_i\|^2$ minimal ist

Bestimmung einer Translation-Matrix

- Gegeben:

  Menge an Wortpaaren (im Wörterbuch) $\{x_i, z_i\}_{i=1}^{n}$

  2 Vektorrepräsentationen pro Paar $x_i \in \mathbb{R}^{d_1}, z_i \in \mathbb{R}^{d_2}$

- Gesucht: Transformationsmatrix *W*
  - sodass: $\sum_{i=1}^{n} \|Wx_i - z_i\|^2$ minimal ist
- Vorgehen: Stochastic Gradient Descent

Bestimmung einer Translation-Matrix

- Recap: Stochastic Gradient Descent

Goal: find parameters θ that reduce cost function $J(\theta)$

Algorithm 1 Pseudocode for SGD

Input:
– function $f(x;\theta)$
– training set of inputs $x_1, \ldots, x_n$ and gold outputs $y_1, \ldots, y_n$
– loss function J
while stopping criteria not met **do**
    Sample a training example x_i, y_i
    Compute the loss $J(f(x_i;\theta), y_i)$
    $\nabla \leftarrow$ gradients of $J(f(x_i;\theta), y_i)$ w.r.t. θ
    Update $\theta \leftarrow \theta - \alpha\nabla$
end while
return θ

Bestimmung einer Translation-Matrix

- Wie kommen wir jetzt ,,vom x zum z”?

$$x_i \in \mathbb{R}^{d_1}, z_i \in \mathbb{R}^{d_2}$$

Bestimmung einer Translation-Matrix

- Wie kommen wir jetzt ,,vom x zum z”?

$$x_i \in \mathbb{R}^{d_1}, z_i \in \mathbb{R}^{d_2}$$

- $z = W \cdot x$

Bestimmung einer Translation-Matrix

- Wie kommen wir jetzt ,,vom x zum z”?

  $$x_i \in \mathbb{R}^{d_1}, z_i \in \mathbb{R}^{d_2}$$

- $z = W \cdot x$

- Geht das nicht sophistizierter?

  → Funktioniert besser als Nearest Neighbour und genauso gut wie neuronale Klassifizierer!

Einsprachige Korpora

WMT11 Dataset

- Shared translation task

Einsprachige Korpora

WMT11 Dataset

- Shared translation task

Preprocessing

- Tokenisieren
- Duplikate entfernen
- Zahlen in einzelne Tokens umschreiben
- Sonderzeichen entfernen
- Phrasen bilden
- Named entities entfernen

Einsprachige Korpora

Language	Training tokens	Vocabulary size
English	575M	127K
Spanish	84M	107K
Czech	155M	505K

Vokabular:

- Wörter, die mindestens 5x im Korpus vorkamen

Dictionaries

Google Translate (GT)

- Ground Truth
- Übersetze häufige Wörter
- Übersetzungen sind nicht immer im Ziel-Vokabular
  - Bei Precision-Berechnung: Unbekannte Worte ignorieren
  - → berichte **Coverage**

Training- und Test-Daten

Training-Daten

- Die 5000 häufigsten Wörter aus einsprachigem Korpus

+

- GT-Übersetzungen

Test-Daten

- Die nächsten 1000 häufigsten Wörter

+

- GT-Übersetzungen

Baselines

1. Baseline

- Längste gemeinsame Zeichenkette
- Implementierung: Edit distance

2. Baseline

- Gleiche Methode wie Mikolov et al., aber mit
- Zählbasierten Vektoren: Kookkurrenzen
  - Dimensionen: gesamtes Vokabular
- Translation Matrix lernen
- In der Ziel-Sprache: Ähnlichstes Wort ist die Prediction

Ergebnisse: WMT11

Translation	Edit Distance		Word Co-occurrence		Translation Matrix		ED + TM		Coverage
	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	
En → Sp	13%	24%	19%	30%	33%	51%	43%	60%	92.9%
Sp → En	18%	27%	20%	30%	35%	52%	44%	62%	92.9%
En → Cz	5%	9%	9%	17%	27%	47%	29%	50%	90.5%
Cz → En	7%	11%	11%	20%	23%	42%	25%	45%	90.5%

Ergebnisse: WMT11

Translation	Edit Distance		Word Co-occurrence		Translation Matrix		ED + TM		Coverage
	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	
En → Sp	13%	24%	19%	30%	33%	51%	43%	60%	92.9%
Sp → En	18%	27%	20%	30%	35%	52%	44%	62%	92.9%
En → Cz	5%	9%	9%	17%	27%	47%	29%	50%	90.5%
Cz → En	7%	11%	11%	20%	23%	42%	25%	45%	90.5%

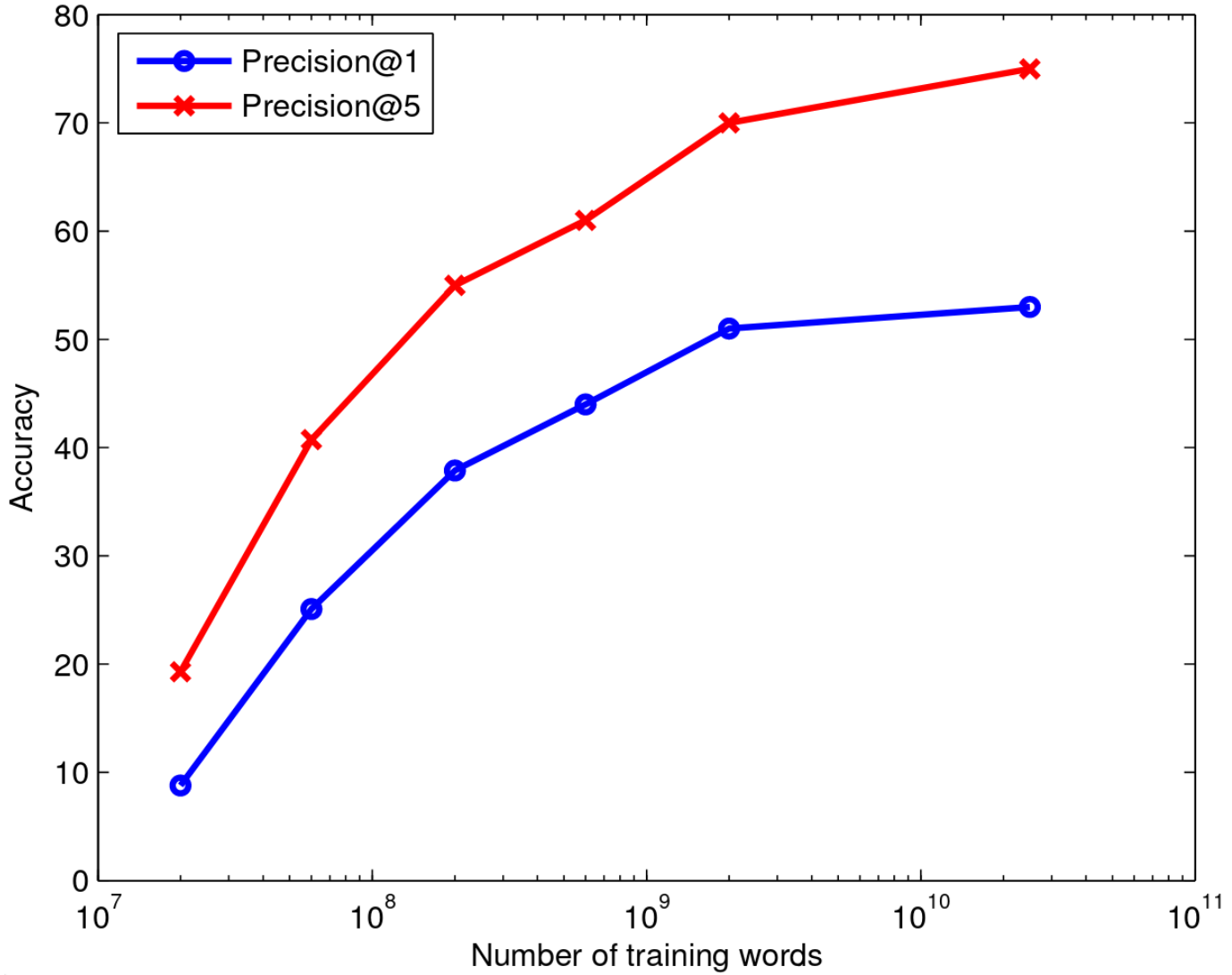
- **Precision at 1**
  - Erfolg, wenn Top-Kandidat die richtige Übersetzung ist
- **Precision at 5**
  - Erfolg, wenn einer der Top 5 Kandidaten die richtige Übersetzung ist

Ergebnisse: Größere Datenmengen

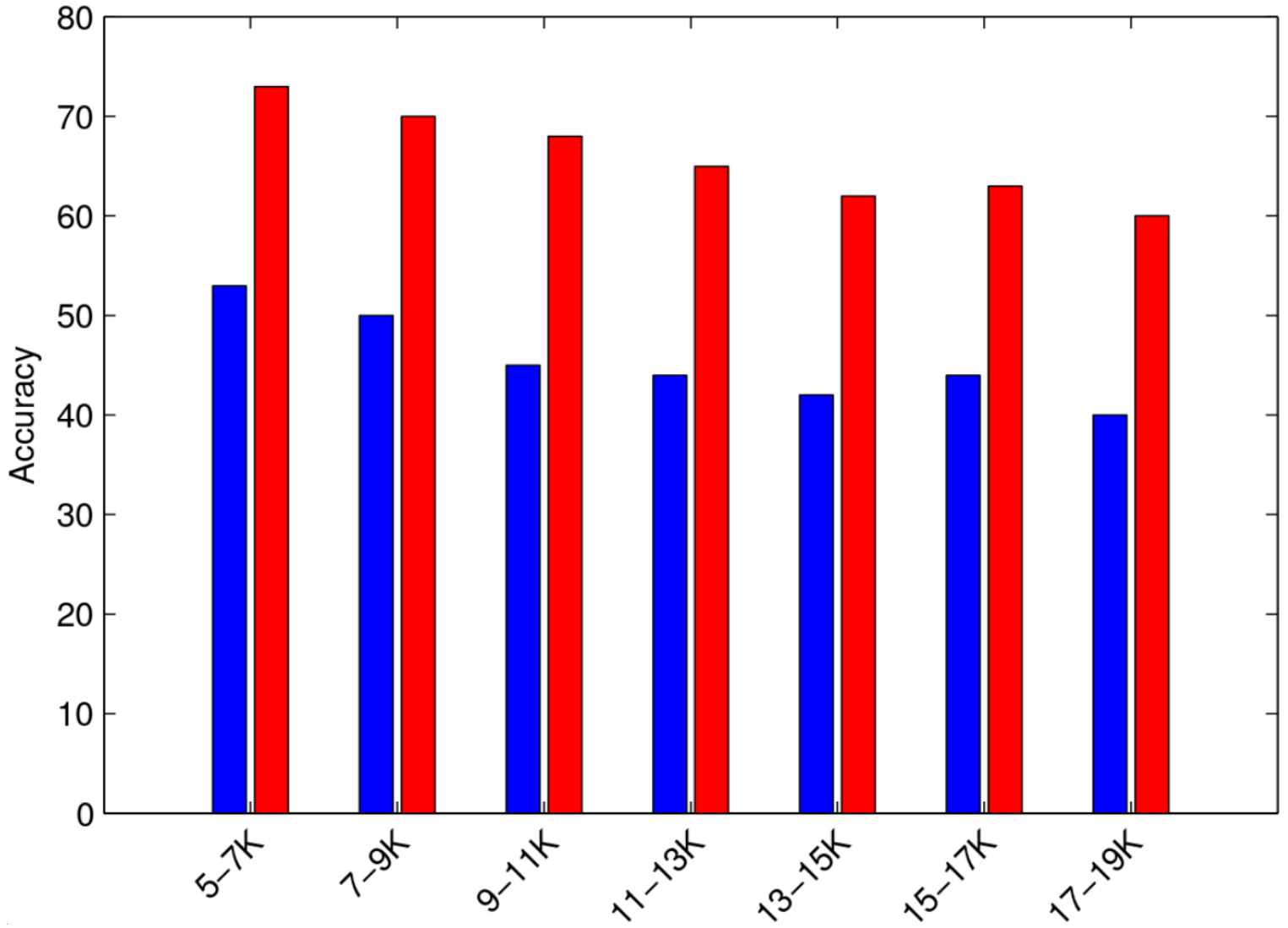
Google News Datensets

- Englisch und Spanisch
- Gleiches Preprocessing
- Training-Set: häufigste 5000 Wörter
- Test-Set: nächsten 1000 Wörter

Ergebnisse: Größere Datenmengen



Ergebnisse: Größere Datenmengen



Distanz als Konfidenzmaß

- Was, wenn wir mehr an **Accuracy** als an Coverage interessiert sind?

Distanz als Konfidenzmaß

- Was, wenn wir mehr an **Accuracy** als an Coverage interessiert sind?
- → Benutze Cosinus-Distanz im Vektorraum als **Konfidenzmaß**

Distanz als Konfidenzmaß

- Was, wenn wir mehr an **Accuracy** als an Coverage interessiert sind?
- → Benutze Cosinus-Distanz im Vektorraum als **Konfidenzmaß**

$$\max_{i \in V} \cos(Wx, z_i)$$

Distanz als Konfidenzmaß

- Was, wenn wir mehr an **Accuracy** als an Coverage interessiert sind?
- → Benutze Cosinus-Distanz im Vektorraum als **Konfidenzmaß**

$$\max_{i \in V} \cos(Wx, z_i)$$

- Suche: das z_i, für das $\cos$ maximal ist

Distanz als Konfidenzmaß

- Was, wenn wir mehr an **Accuracy** als an Coverage interessiert sind?
- → Benutze Cosinus-Distanz im Vektorraum als **Konfidenzmaß**

$$\max_{i \in V} \cos(Wx, z_i)$$

- Suche: das z_i, für das $\cos$ maximal ist
- Wenn Konfidenz kleiner als festgelegter Threshold: Überspringen

Distanz als Konfidenzmaß

- Was, wenn wir mehr an **Accuracy** als an Coverage interessiert sind?
- → Benutze Cosinus-Distanz im Vektorraum als **Konfidenzmaß**

$$\max_{i \in V} \cos(Wx, z_i)$$

- Suche: das z_i, für das $\cos$ maximal ist
- Wenn Konfidenz kleiner als festgelegter Threshold: Überspringen

Threshold	Coverage	P@1	P@5
0.0	92.5%	53%	75%
0.5	78.4%	59%	82%
0.6	54.0%	71%	90%
0.7	17.0%	78%	91%

EN → ES

Distanz als Konfidenzmaß

- Was, wenn wir mehr an **Accuracy** als an Coverage interessiert sind?
- → Benutze Cosinus-Distanz im Vektorraum als **Konfidenzmaß**

$$\max_{i \in V} \cos(Wx, z_i)$$

- Suche: das z_i, für das $\cos$ maximal ist
- Wenn Konfidenz kleiner als festgelegter Threshold: Überspringen

Threshold	Coverage	P@1	P@5
0.0	92.5%	53%	75%
0.5	78.4%	59%	82%
0.6	54.0%	71%	90%
0.7	17.0%	78%	91%

EN → ES

Threshold	Coverage	P@1	P@5
0.0	92.5%	58%	77%
0.4	77.6%	66%	84%
0.5	55.0%	75%	91%
0.6	25.3%	85%	93%

EN → ES (mit Edit Distance)

Beispiele: Ein Blick in die Daten

Spanish word	Computed English Translations	Dictionary Entry
emociones	emotions emotion feelings	emotions
protegida	wetland undevelopable protected	protected
imperio	dictatorship imperialism tyranny	empire
determinante	crucial key important	determinant
preparada	prepared ready prepare	prepared
millas	kilometers kilometres miles	miles
hablamos	talking talked talk	talk
destacaron	highlighted emphasized emphasised	highlighted

Table 5: Examples of translations of out-of-dictionary words from Spanish to English. The three most likely translations are shown. The examples were chosen at random from words at ranks 5K–6K. The word representations were trained on the large corpora.

Beispiele: Ein Blick in die Daten

Spanish word	Computed English Translations	Dictionary Entry
emociones	emotions emotion feelings	emotions
protegida	wetland undevelopable protected	protected
imperio	dictatorship imperialism tyranny	empire
determinante	crucial key important	determinant
preparada	prepared ready prepare	prepared
millas	kilometers kilometres miles	miles
hablamos	talking talked talk	talk
destacaron	highlighted emphasized emphasised	highlighted

- Auch falsche Übersetzungen sind semantisch verwandt

Table 5: Examples of translations of out-of-dictionary words from Spanish to English. The three most likely translations are shown. The examples were chosen at random from words at ranks 5K–6K. The word representations were trained on the large corpora.

Beispiele: Ein Blick in die Daten

English word	Computed Spanish Translation	Dictionary Entry
pets	mascotas	mascotas
mines	minas	minas
unacceptable	inaceptable	inaceptable
prayers	oraciones	rezo
shortstop	shortstop	campocorto
interaction	interacción	interacción
ultra	ultra	muy
beneficial	beneficioso	beneficioso
beds	camas	camas
connectivity	conectividad	conectividad
transform	transformar	transformar
motivation	motivación	motivación

- High-Confidence-Translations (Score > 0.5)
- Nutzt auch Edit-Distance

Weiterer Use-Case

- **Was können wir mit den Modellen machen?**

Weiterer Use-Case

- **Was können wir mit den Modellen machen?**
  - Offensichtlichster Nutzen ist natürlich das Anfertigen möglichst guter Übersetzungen zwischen Sprachen, auch wenn nur wenige bilinguale Ressourcen zur Verfügung stehen

Weiterer Use-Case

- **Was können wir mit den Modellen machen?**
  - Offensichtlichster Nutzen ist natürlich das Anfertigen möglichst guter Übersetzungen zwischen Sprachen, auch wenn nur wenige bilinguale Ressourcen zur Verfügung stehen
- **Fallen euch weitere Use-Cases ein?**

Weiterer Use-Case

- **Was können wir mit den Modellen machen?**
  - Offensichtlichster Nutzen ist natürlich das Anfertigen möglichst guter Übersetzungen zwischen Sprachen, auch wenn nur wenige bilinguale Ressourcen zur Verfügung stehen
  - Wir können die Modelle auch nutzen, um Fehler in den Wörterbüchern zu finden

Weiterer Use-Case

- **Was können wir mit den Modellen machen?**
  - Offensichtlichster Nutzen ist natürlich das Anfertigen möglichst guter Übersetzungen zwischen Sprachen, auch wenn nur wenige bilinguale Ressourcen zur Verfügung stehen
  - Wir können die Modelle auch nutzen, um Fehler in den Wörterbüchern zu finden

    → Berechne Distanz zwischen Output-Wort und Übersetzung, die im Wörterbuch steht

    → Wenn Distanz > X: Überprüfe Wörterbucheintrag

Weiterer Use-Case

English word	Computed Czech Translation	Dictionary Entry
said	řekl *(said)*	uvedený *(listed)*
will	může *(can)*	vůle *(testament)*
did	udělal *(did)*	ano *(yes)*
hit	zasáhl *(hit)*	hit -
must	musí *(must)*	mošt *(cider)*
current	stávající *(current)*	proud *(stream)*
shot	vystřelil *(shot)*	shot -
minutes	minut *(minutes)*	zápis *(enrollment)*
latest	nejnovější *(newest)*	poslední *(last)*
blacks	černoši *(black people)*	černá *(black color)*
hub	centrum *(center)*	hub -

Weiterer Use-Case

English word	Computed Czech Translation	Dictionary Entry
said	řekl *(said)*	uvedený *(listed)*
will	může *(can)*	vůle *(testament)*
did	udělal *(did)*	ano *(yes)*
hit	zasáhl *(hit)*	hit -
must	musí *(must)*	mošt *(cider)*
current	stávající *(current)*	proud *(stream)*
shot	vystřelil *(shot)*	shot -
minutes	minut *(minutes)*	zápis *(enrollment)*
latest	nejnovější *(newest)*	poslední *(last)*
blacks	černoši *(black people)*	černá *(black color)*
hub	centrum *(center)*	hub -

,,We chose the examples manually, so this demonstration is highly subjective.”

Was, wenn Wort nicht gleich Wort ist?

- Im Vietnamesischen ist das Konzept von Wörtern nicht wie im Englischen oder Deutschen

Was, wenn Wort nicht gleich Wort ist?

- Im Vietnamesischen ist das Konzept von Wörtern nicht wie im Englischen oder Deutschen
- Beispiel:

  Auto = **xe hơi** (Wagen Luft)

Was, wenn Wort nicht gleich Wort ist?

- Im Vietnamesischen ist das Konzept von Wörtern nicht wie im Englischen oder Deutschen
- **Wie könnte man damit umgehen?**

Was, wenn Wort nicht gleich Wort ist?

- Im Vietnamesischen ist das Konzept von Wörtern nicht wie im Englischen oder Deutschen
- Idee: Man betrachtet bei solchen Sprachen keine Wörter, sondern Phrasen, die dann Wörtern oder kurzen Phrasen der Zielsprache entsprechen

Was, wenn Wort nicht gleich Wort ist?

- Im Vietnamesischen ist das Konzept von Wörtern nicht wie im Englischen oder Deutschen
- Idee: Man betrachtet bei solchen Sprachen keine Wörter, sondern Phrasen, die dann Wörtern oder kurzen Phrasen der Zielsprache entsprechen

Table 8: The accuracy of our translation method between English and Vietnamese. The edit distance technique did not provide significant improvements. Although the accuracy seems low for the EN→VN direction, this is in part due to the large number of synonyms in the VN model.

Threshold	Coverage	P@1	P@5
En → Vn	87.8%	10%	30%
Vn → En	87.8%	24%	40%

Fazit

- Großes Potential, dem Problem kleiner bilingualer Ressourcen zu begegnen

Fazit

- Großes Potential, dem Problem kleiner bilingualer Ressourcen zu begegnen
- Nicht nur zur Erweiterung, sondern auch zur Verbesserung vorhandener Wörterbücher geeignet

Fazit

- Großes Potential, dem Problem kleiner bilingualer Ressourcen zu begegnen
- Nicht nur zur Erweiterung, sondern auch zur Verbesserung vorhandener Wörterbücher geeignet
- Funktioniert auch für wenig verwandte Sprachen (EN-CZ) recht gut

Fazit

- Großes Potential, dem Problem kleiner bilingualer Ressourcen zu begegnen
- Nicht nur zur Erweiterung, sondern auch zur Verbesserung vorhandener Wörterbücher geeignet
- Funktioniert auch für wenig verwandte Sprachen (EN-CZ) recht gut
- Unterschiedliche Wortkonzepte sind problematisch, betreffen aber nur wenige Sprachen und man kann statt Wörtern auch Phrasen betrachten

Quellen

- **Auf den Folien angegeben**
- **Sonst:**

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Vielen Dank für Ihre und eure Aufmerksamkeit!